Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

CLUSTERING CHROLOPHYLL-A SATELLITE DATA USING QUANTILES

By †Carlo Gaetan, †Paolo Girardi*, †Roberto Pastres and ‡Antoine Mangin

†*DAIS, Università Ca' Foscari - Venezia, Italy*

‡*ACRI-ST, Sophia-Antipolis, France*

The use of water quality indicators is of crucial importance to identify risks to the environment, society and human health. In particular, the Chlorophyll type A (Chl-a) is a shared indicator of trophic status and for monitoring activities it may be useful to discover local dangerous behaviours (for example the anoxic events). In this paper we consider a comprehensive data set, covering the whole Adriatic Sea, derived from Ocean Colour satellite data, during the period 2002-2012 with the aim of identifying homogeneous areas. Such zonation is becoming extremely relevant for the implementation of European policies, such the Marine Strategy Framework Directive. As an alternative to clustering based on an "average" value over the whole period, we propose a new clustering procedure for the time series. The procedure shares some similarities with the functional data clustering and combines nonparametric quantile regression with an agglomerative clustering algorithm. This approach permits to take into account some features of the time series as non stationarity in the marginal distribution and the presence of missing data. A small simulation study is also presented for illustrating the relative merits of the procedure.

1. INTRODUCTION

The amount of datasets produced and made available through satellite imaging has rapidly increased in the last few years. These datasets, when collected over times, provide a rich source of information on the dynamic nature of earth surface processes and can be used for monitoring biological, climate and ecological dynamics.

*Address for correspondence: Paolo Girardi, DAIS, Università Ca' Foscari - Venezia, Scientific Campus, via Torino 155, 30170 Venezia Mestre, ITALY. E-mail: paolo.girardi@unive.it.

Keywords and phrases: Functional data clustering, Quantile sheet, Non parametric regression, Clustering methods, Surface water classification, Satellite data

In this paper we are interested in the temporal structure of re-processed satellite data concerning the concentration of Chlorophyll type-a (Chl-a), in the superficial water of the Adriatic Sea. Chl-a is an indicator of the biomass of phytoplankton (e.g. photosynthetic algae, from those unicellular to multicellular ones, Huot et al. (2007); Behrenfeld and Falkowski (1997)). The level of Chl-a shows seasonal variation, due to inter-annual changes of water temperature, available light, and, to some extent, inorganic nutrients, such as dissolved nitrogen, phosphorus and silicate (Yoder et al., 1993). Furthermore, nutrient enrichment may cause eutrophic conditions in which marked algal blooms may be followed by nutrient depletion and a rapid decrease in algal biomass. The subsequent degradation of the algal biomass may then lead to hypoxic or even anoxic events. The European WFD 2000/60/EC defines a series of threshold values for the concentration in order to define areas at "high impact" of Chlorophyll type-a concentrations. Then areas have been regrouped only considering the average value over a period, usually one year.

In the last years, many studies have used satellite data of the Chl-a concentration in order to verify the spatial distribution of water eutrophication in the Adriatic Sea (Giani et al., 2012; Djakovac et al., 2012; Marini et al., 2010; Mélin et al., 2011). Satellite data provides valuable temporal information over a grid. In our context the clustering of the time series observed at each point of the grid could help to define areas of "high impact", which could be characterized by high peaks of Chl-a, and also discover homogeneous areas using the dynamics related to the Chl-a concentration. This type of clustering procedure is likely to provide more reliable results in comparison to simpler classifications, based on the average value of the Chl-a over the time, since it takes into account both the seasonal and inter-annual variability.

In order to extract useful information from these satellite image time series, we need a new clustering procedure that should be able to address some issues. First the time series over a given site can become irregular in terms of temporal sampling due to meteorological phenomena, such as clouds. Secondly the phenomenon of interest exhibits a periodic behaviour which can be slightly modulated by climate and/or anthropic artifacts. These modulations result in distortions of canonical temporal profile (Petitjean, Inglada and Gançarski, 2012), which should be stable if not affected by climate changes or anthropic influences. Therefore the resulting time series appears irregularly sampled, non-stationary in mean and with different variability in time. These features make the

clustering task quite challenging.

Several approaches to time series clustering have been proposed in the literature (see Liao, 2005, for a recent review). The simplest procedures treat time series as multivariate data and cluster them using some dissimilarity measures concerning observations or sample summaries, such as mean, median or standard deviation computed from the raw data. This approach shows some limits: the results of the classification are very sensitive to the choice of a given dissimilarity measure and these methods do not take into account the "structural features" of the time series, such as local or global trends and isolated peaks. Wang, Smith and Hyndman (2006) try an approach for overcoming these limits by considering the structural characteristics of a time series.

In the model-based approach, each time series is thought to be generated by some kind of model, which is generally assumed to be a linear stationary process. Therefore, dissimilarity measures (see Piccolo, 1990, for instance) in the space of model parameters have to be introduced in order to perform a classification. In such approach the mechanism assigning each time series to a particular group is deterministic. A related approach (see Frühwirth-Schnatter and Kaufmann, 2008, for example) considers that this mechanism can be captured by a latent variable that drives the assignment. The distribution of this latent variable may be either independent of the specific time series or may depend on time series characteristics.

If the time series are irregularly sampled another possibility is considering a time series, as a discrete and noised representation of a curve $\mu(t)$,

$$Y(t) = \mu(t) + \varepsilon(t)$$

where t indicates the time and $\varepsilon(t)$ is a random error with zero mean. This representation leads to a functional representation of the data (Ramsay and Silverman, 2005). The curve is interpreted as a mean function, i.e. $E(Y(t)) = \mu(t)$ which is determined as a finite expansion of appropriate basis. Subsequently, a clustering procedure is applied to the basis expansion coefficients (Abraham et al., 2003; Antoniadis et al., 2013; Jacques and Preda, 2014; Haggarty, Miller and Scott, 2015). As an alternative working hypothesis, each time series is supposed to be generated by a mixture of underlying distributions (James and Sugar, 2003; Pastres, Pastore and Tonellato, 2011; Haggarty et al., 2012; Nieto-Barajas and Contreras-Cristán, 2014).

Clustering techniques for functional data are mainly concentrated on the mean function and treat

higher moments as constant nuisance parameters. As a consequence, clustering crucially relies on assumptions such as homoscedasticity and symmetry of the marginal distribution of the time series.

For these reasons, we introduce a new method for clustering time series taking into account more properly the variability. The novelty lies in the use of a technique based on nonparametric quantile regressions with the scope of estimating the temporal distribution patterns (Cheng, 1983; Koenker, Ng and Portnoy, 1994). More precisely, by varying the quantiles of interest in the nonparametric quantile regression, we are able to make inferences about the marginal distribution of the data at each time. However care has to be taken in order to avoid the problem that curves pertaining to different quantile regressions may cross, as might happen when non stationary time series exhibits high variability. To avoid this problem, we estimated a "quantile sheet" (Schnabel and Eilers, 2013) that allows the simultaneous estimation of all quantile curves without intersections. Subsequently, an agglomerative clustering algorithm is used to group the basis coefficients using an opportune distance measure (e.g. L2-norm). Therefore, our methodology shares some similarity with clustering functional data but, due to its flexibility, it allows one to cope with different variability in time.

The paper is organized as follows. The next section illustrates our data on Chl-a that we use in the subsequent analysis. In Section 3 we introduce the quantile sheet and its estimation and in Section 4 we present the clustering procedure. Section 5 is devoted to a small simulation study for illustrating the performances of the procedure. Section 6 considers the application of the procedure to the Chrolophyll-a satellite data. In the last section we discuss the relative strengths and weaknesses of our proposal.

2. DATA

We have considered the monthly mean values of the Chl-a concentration in the Adriatic Sea from January 2002 to December 2012. Data were obtained by calibrating Ocean Colour data provided by different satellite missions, such as MERIS, SeaWiFS and MODIS, and the data set was made available by ACRI (`http://hermes.acri.fr`) in the framework of the GlobColour Project (Maritorena et al., 2010).

For each month, we have extracted gridded data with a resolution of 192×240 points (longitude: 12.02^oE to 21.98^oE; latitude: 38.02^oN to 45.98^oN; 4Km scale). Satellite data are usually affected by a measurement error due to many physical factors. Along the coasts, the low bathimetry and

the tide level affect the quality of the measurement and lead to lack of many data. Therefore we have analysed a coarse resolution of 8km resulting in a grid of 96×120 points where $2,168$ points cover the Adriatic Sea (Figure 1-a). Satellite Chl-a data can be approximated with a log-normal distribution (Campbell, 1995) and we have considerered in our analysis log-transformed values.

Chl-a concentration is subjected to seasonal and annual changes. It is usually influenced by rivers and their input of nutrients, climatic conditions (sunlight, rain, temperature, etc..), water depth and hydrodynamics. As an example, in Figure 1(b-c-d) we show the temporal evolution of the logarithm of the Chl-a concentration in three sites representative of three areas within the studied area. The patterns are quite different, being characterized by different seasonal behaviours: in general we observe higher spring values and lower summer ones associated to a different grade of heteroscedasticity at sites 1 and 2. Furthermore, sites (2 and 3) located at the same distance from the coast exhibit a very different trend.

3. QUANTILE SHEET AND ITS ESTIMATION

We look at time series of data collected on a site, i.e. $y_i = \{y_i(t_j), \quad t_j \in \mathcal{T}_i\}$ where $y_i(t_j)$ represents the outcome of a random variable $Y_i(t_j)$ that we observe at the time instants t_j on the site i of coordinates s_i. The set $\mathcal{T}_i = \{t_1, \ldots, t_{J_i}\}$ denotes the set of time instants in which measurements are taken at site i. In the following a crucial assumption is that the whole temporal structure of y_i is described in terms of the marginal distributions of $Y_i(t)$. In order to define the quantile sheet of a time series we concentrate on a single time series and for simplicity we drop the index i.

The τ-quantile function, $g(t, \tau)$, of $Y(t)$ is a function of t such that

$$Pr\left(Y(t) \leq g(t, \tau)\right) = \tau, \qquad 0 \leq \tau \leq 1.$$

For a while we fix a value for the probability τ and we assume $g(t, \tau)$ to be a smooth function in t that belongs to an appropriately chosen function space. An estimate of $g(t, \tau)$ is obtained (Koenker, 2005) by minimizing the fitness function

$$\sum_{j=1}^{J} u_j(\tau)|y(t_j) - g(t_j, \tau)|$$

where $u_j(\tau) = \tau 1_{\{y(t_j) > g(t_j,\tau)\}} + (1 - \tau)1_{\{y(t_j) \leq g(t_j,\tau)\}}$.

The minimization process is difficult to perform because the convergence may take long time, especially in presence of a large dataset and of many quantiles. Here for computational reasons, we adopt an iterative algorithm that is a variant of the Majorization-Minimization (MM) algorithm following Schlossmacher (1973) and Schnabel and Eilers (2013).

We use a quadratic fitness function

$$\sum_{j=1}^{J} w_j(\tau)(y(t_j) - g(t_j, \tau))^2$$

where $w_j(\tau) = u_j(\tau)/\sqrt{\{y(t_j) - g(t_j,\tau)\}^2 + \beta^2}$. The value of β has to be carefully selected. Schnabel and Eilers (2013) recommended a small number of the order of 10^{-4} times the maximum absolute value of values $u_j(\tau)$.

Finally we suppose that we can represent g as a linear combination of L B-spline basis functions $B_l(t)$, $l = 1, \ldots, L$, namely

$$g(t, \tau) = \sum_{l=1}^{L} \gamma_l B_l(t). \tag{1}$$

Using a penalized regression spline approach (Eilers and Marx, 1996), we can derive an estimate of $g(t,\tau)$, $\widehat{g}(t,\tau) = \sum_{l=1}^{L} \widehat{\gamma}_l B_l(t)$, solving

$$\widehat{\gamma} = \arg\min_{\gamma} (y - B\gamma)^\top W (y - B\gamma) + \gamma^\top (\lambda\Omega)\gamma, \qquad \lambda > 0$$

where $y = (y(t_1), \ldots, y(t_J))^\top$, $\gamma = (\gamma_1, \ldots, \gamma_L)^\top$, B is a J by L matrix with elements $b_{jl} = B_l(t_j)$ and W is a J by J diagonal matrix with diagonal element $w_{jj} = w_j(\tau)$.

The penalty matrix $\lambda\Omega$ term is introduced to penalize the roughness of the fitted τ-quantile function $\widehat{g}(t,\tau)$ and λ is a penalty parameter. For evenly spaced values of t_j, Ω can be chosen such that $\gamma^\top \Omega \gamma = \sum_{l=2}^{L-1} (\gamma_{l+1} - 2\gamma_l + \gamma_{l-1})^2$ is the squared second difference penalty, namely $\Omega = D^\top D$ and

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \ldots \\ 0 & 1 & -2 & 1 & \ldots \\ 0 & 0 & 1 & -2 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Fixing the weights $w_j(\tau)$ and the penalty parameter λ we have an explicit expression of $\widehat{\gamma}$, i.e.

$$\widehat{\gamma} = (B^\top W B + \lambda\Omega)^{-1} B^\top W y. \tag{2}$$

Therefore we can devise an iterative algorithm to calculate the penalized spline estimate of the $\tau-$quantile function by alternating between evaluating (2), the residuals and recomputing the weights $w_j(\tau)$ as $u_j(\tau)/\sqrt{\{y(t_j) - \widehat{g}(t_j,\tau)\}^2 + \beta^2}$ until convergence. A convenient starting point for the weights is $w_j(\tau) = 1$ for all j. The convergence of the MM algorithm follows from the general result in Hunter and Lange (2000).

Varying t and the probability τ we can obtain a complete description of the marginal distributions of $Y(t)$. How we estimate in this case the function $g(t,\tau)$ is less straightforward. In principle we can fit separate models for a grid of probabilities τ, however curves pertaining to different quantiles may cross. Solutions to this problem can be obtained by combining all quantile fits in one joint model based on, for example, location–scale models (He, 1997) or by estimating simultaneously all parameters for each quantile curve (Bondell, Reich and Wang, 2010).

Schnabel and Eilers (2013) interpreted $g(t,\tau)$ as a surface, called a *quantile sheet*. A parsimonious representation of $g(t,\tau)$ is obtained by considering a B-spline basis for the probability τ and a tensor product of the basis for the times and the probability. More formally, let $\tilde{B}_1(\tau),\ldots,\tilde{B}_M(\tau)$ be the B-spline basis for τ, and we suppose that

$$g(t,\tau) = \sum_{l=1}^{L}\sum_{m=1}^{M} \gamma_{lm} B_l(t)\tilde{B}_m(\tau). \tag{3}$$

Fixing K values $\tau_1,\ldots,\tau_K$, we estimate γ_*, the vector of the unknown coefficients γ_{lm} by

$$\begin{aligned} \widehat{\gamma}_* &= \text{argmin}_{\gamma_*} \sum_{j=1}^{J}\sum_{k=1}^{K} w_j(\tau_k)\left(y(t_j) - \sum_{l=1}^{L}\sum_{m=1}^{M}\gamma_{ll}B_l(t_j)\tilde{B}_m(\tau_k)\right)^2 + \gamma_*^\top \Omega_* \gamma_* \\ &= \text{argmin}_{\gamma_*}\ (y_* - B_*\gamma_*)^\top W_*(y_* - B_*\gamma_*) + \gamma_*^\top\Omega_*\gamma_* \qquad (4) \end{aligned}$$

Here $y_* = 1_K^\top \otimes y^\top$, $W_* = \text{diag}(vec(W))$, where W is a J by K matrix with elements $w_{jk} = w_j(\tau_k)$, $B_* = \tilde{B}\otimes B$ and $\tilde{B}$ is a K by M matrix with elements $b_{km} = \tilde{B}_m(\tau_k)$. The penalty matrix Ω_* is

$$\Omega_* = \lambda(I_M \otimes \Omega) + \tilde{\lambda}(\tilde{\Omega}\otimes I_L), \qquad \lambda,\tilde{\lambda} > 0$$

where I_M and I_L are two identity matrices of size M and L, while and Ω and $\tilde{\Omega}$ are the squared second difference penalty for times and probability, respectively.

The formula in (4) gives us a solution similar to (2), namely

$$\widehat{\gamma}_* = (B_*^\top W_* B_* + \Omega_*)^{-1} B_*^\top W_* y_*. \tag{5}$$

However, the Kronecker product in formula (5) is very memory-demanding and computationally cumbersome when we use a high-level language like Matlab or R (R Core Team, 2014), but in our R implementation we rearrange the computations as in Eilers, Currie and Durbán (2006).

The full implementation requires to fix the amount of smoothing to be fixed. Schnabel and Eilers (2013) suggested to perform n-fold cross-validation for choosing the penalty parameters λ and $\tilde{\lambda}$ but implementing this method inside the clustering procedure requires a lot of computational resources. Differently we have preferred to adopt a generalized approximate cross-validation (GACV) criterion (Yuan, 2006), namely

$$GACV(\lambda, \tilde{\lambda}) = \frac{(y_* - B_*\hat{\gamma}_*)^\top W_*(y_* - B_*\hat{\gamma}_*)}{n \times m - \mathrm{tr}(H_*)} \tag{6}$$

where $H_* = B_*(B_*^\top W_* B_* + \Omega_*)^{-1} B_*^\top W_*$. The evaluation of the trace is greatly simplified if we consider $\mathrm{tr}(B_*(B_*^\top W_* B_* + \Omega_*)^{-1} B_*^\top W_*) = \mathrm{tr}((B_*^\top W_* B_* + \Omega_*)^{-1} B_*^\top W_* B_*)$.

In the clustering procedure a vector $\gamma_{*,i}$ has to be estimated for each time series and in principle we can suppose each quantile sheet has different amount of smoothing. However we have experimented that this strategy yields to a less robust clustering procedure so we prefer to choose the amount of smoothing by minimizing the overall criterion $\sum_{i=1}^{N} GACV_i(\lambda, \tilde{\lambda})$ where $GACV_i$ is of the form as (6) evaluated for the $i-$th time series.

Model (3) can be extended with regard to seasonal variability. Judging from the observed monthly values in Figure 1 we can see a strong and varying seasonal pattern. For this reason we propose a modulation model (Eilers et al., 2008), namely

$$g(t, \tau) = \tilde{g}_1(t, \tau) + \tilde{g}_2(t, \tau)\cos(\pi/6\, t) + \tilde{g}_3(t, \tau)\sin(\pi/6\, t). \tag{7}$$

For a fixed probability τ, g_1 accounts for a smooth temporal trend while g_2 and g_3 are smooth functions that describe the local amplitudes of the cosine and sine waves. Again, we use B-spline functions as in (3) for specifying $\tilde{g}_i$, $i = 1, 2, 3$, that is

$$\tilde{g}_i(t, \tau) = \sum_{l=1}^{L_i} \sum_{m=1}^{M} \gamma_{lm,i} B_l(t) \tilde{B}_m(\tau).$$

Setting $B_l^C(t) = \cos(\pi/6\, t) B_l(t)$, and $B_l^S(t) = \sin(\pi/6\, t) B_l(t)$ we can rewrite the model (7) as

$$\tilde{g}_i(t, \tau) = \sum_{l=1}^{L_1} \sum_{m=1}^{M} \gamma_{lm,1} B_l(t) \tilde{B}_m(\tau) + \sum_{l=1}^{L_2} \sum_{m=1}^{M} \gamma_{lm,2} B_l^C(t) \tilde{B}_m(\tau) + \sum_{l=1}^{L_3} \sum_{m=1}^{M} \gamma_{lm,3} B_l^S(t) \tilde{B}_m(\tau). \tag{8}$$

Stacking the coefficients $\gamma_{lm,i}$ into the vector γ_* the estimate $\hat{\gamma}_*$ can be derived as in formula (4) provided that the matrix B is replaced with the matrix $[B, CB, SB]$, where $C = \text{diag}\{\cos(\pi/6t_i)\}$, and $S = \text{diag}\{\sin(\pi/6t_i)\}$ and the matrix $\lambda(I_M \otimes \Omega)$ is replaced by the block diagonal matrix $I_M \otimes \text{diag}(\lambda_1, \lambda_2, \lambda_3) \otimes \Omega$. The parameters $\lambda_i > 0$, $i = 1, 2, 3$ control the amount of smoothing of the overall trend and the modulated seasonal components.

4. CLUSTERING TIME SERIES BY MEANS OF THE QUANTILE SHEETS

For clustering I time series y_i, $i = 1, \ldots, I$, we use a partition around medoid (PAM) algorithm (Kaufman and Rousseeuw, 1990) coupled with Gap Statistic (Tibshirani, Walther and Hastie, 2001).

We define the similarity of two time series y_i and $y_{i'}$ in terms of the distance between their quantile sheets according the L^2 distance

$$d_{ii'} = d(g_i, g_{i'}) = \sqrt{\int\int (g_i(t;\tau) - g_{i'}(t;\tau))^2 dt d\tau}.$$

This definition allows us to take into account shifts not only in the mean, but also in the marginal distribution during the whole period. The evaluation of this distance is simplified in the representation (3) because each quantile sheet g_i can be summarized by the $L \times M$ vector, $\gamma_{*,i}$, of its coefficients.

It turns out that

$$\begin{aligned} d^2_{ii'} &= \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{l'=1}^{L}\sum_{m'=1}^{M} (\gamma_{lm,i} - \gamma_{lm,i'}) v_{ll'} \tilde{v}_{mm'} (\gamma_{l'm',i} - \gamma_{l'm',i'}) \\ &= (\gamma_{*,i} - \gamma_{*,i'})^\top (\tilde{V} \otimes V)(\gamma_{*,i} - \gamma_{*,i'}) \end{aligned}$$

with $V = [v_{ll'}]$, $v_{ll'} = \int B_l(t) B_{l'}(t) dt$, and $\tilde{V} = [\tilde{v}_{mm'}]$, $\tilde{v}_{mm'} = \int \tilde{B}_m(\tau) \tilde{B}_{m'}(\tau) d\tau$. Finally the resulting dissimilarity matrix is the input matrix for the PAM algorithm.

The number of clusters is chosen by means of the Gap Statistic (Tibshirani, Walther and Hastie, 2001) as implemented in the function `clusGap` of the R package *cluster*. We have clustered I time series into N clusters $C_1, C_2, \ldots, C_N$ with C_n denoting the indices of the times series in cluster n; the Gap Statistic is based on the within-cluster sum of squared distances from the cluster means

$$W(N) = \sum_{n=1}^{N} \frac{1}{2|C_n|} D_n$$

with $D_n = \sum_{i,i' \in C_n} d_{ii'}$ and $|C_n|$ the cardinality of C_n. The Gap for N clusters is defined as

$$Gap(N) = E[\log(W^*(N))] - \log(W(N))$$

where $E[\log(W^*(N))]$ is the expectation of the logarithm of the within-cluster variation we would see if we instead had points distributed uniformly over an encapsulating box. The quantity $E[\log(W_N^*)]$ is computed by simulation and we average the log within-cluster variation over B simulated uniform data sets. Finally the proper number of clusters for the given data set is the smallest N such that

$$Gap(N) \geq Gap(N+1) - s_{N+1}$$

where $s_N = \sqrt{1+1/B}\, sd_N$ is the simulation error calculated from the standard deviation sd_N of the B replicates (Tibshirani, Walther and Hastie, 2001).

In our motivating example time series are geographically referenced samples and the spatial dependence has to be taken into account. A possible solution is adjusting the dissimilarity matrix following the idea in Giraldo, Delicado and Mateu (2012). We consider the I estimated quantile sheets, $\widehat{g}_i$, as georeferenced functional data and we define a new similarity measure for the site s_i and $s_{i'}$ as $d_{ii'}^S = d(\widehat{g}_i, \widehat{g}_{i'})\Delta(s_i, s_{i'})$, where

$$\Delta(s_i, s_{i'}) = \frac{1}{2}\int\int \mathbb{E}\left[\widehat{g}_i(t,\tau) - \widehat{g}_j(t,\tau)\right]^2 dt\; d\tau.$$

Assuming that $\{\hat{g}_i\}$ is a spatial intrinsic stationary functional random field, the function $\Gamma(s_i - s_{i'}) = \Delta(s_i, s_{i'})$ is a valid semi-variogram that can be estimated by using the moment estimator

$$\widehat{\Gamma}(h) = \frac{1}{2|N(h)|}\sum_{i,i'\in|N(h)|} d^2(\widehat{g}_i, \widehat{g}_{i'}) = \frac{1}{2|N(h)|}\sum_{i,i'\in N(h)}\int\int (\widehat{g}_i(t;\tau) - \widehat{g}_{i'}(t;\tau))^2 dt\; d\tau \tag{9}$$

where $|N(h)|$ is the number of element of the bin $N(h)$, i.e. a subset of possible lags $s_i - s_{i'}$ and h a representative member of $N(h)$.

Once we have estimated $\Gamma(h)$ for a sequence of values h a parametric model $\Gamma(h;\theta)$ is fitted by weighted least squares (Cressie, 1993, p. 95). Finally we use

$$\widehat{d}_{ii'}^S = d(\widehat{g}_i, \widehat{g}_{i'})\Gamma(s_i - s_{i'}; \hat{\theta}) \tag{10}$$

as element of the dissimilarity matrix for the PAM algorithm.

The assumption of spatial stationarity seems too restrictive in our motivating example. Owing to the representation (3) of the quantile sheet, spatial non-stationarity can arise if the distribution of the estimated coefficients $\hat{\gamma}_{*,i}$ depends on s_i. A solution that we adopted is fitting a trend surface on the values $\hat{\gamma}_{lm,i}$ and inserting the detrended coefficients into (9).

5. SIMULATED DATA

In this small simulation study we are mainly interested in seeing how our proposal can cluster different temporal patterns that we have encountered in our dataset. For simplicity we have considered non stationary time series which are spatially independent.

We have considered two examples and in each example we have simulated datasets structured in 3 clusters. Each cluster contains one hundred time series generated from a non-stationary model. The number of observations for each time series is 132. The clustering procedure based on quantile sheet (QSC) in the sequel, has been compared with four other clustering procedures, namely

1. competitor clustering procedure based on quantiles (CQC), according to which a set of conditional quantile splines curves are fitted to the time series and then the estimated coefficients are clustered using the PAM algorithm.
2. functional curve clustering (FCC) procedure in which for each time series first we estimate a mean curve using regression splines (Abraham et al., 2003). Then the estimated spline coefficients are partitioned using the PAM algorithm;
3. wavelet-based clustering (WAC) procedure, see Antoniadis et al. (2013) for details, in which we extract time-series features using the wavelet coefficients of an orthonormal basis and calculating the distribution of energy across scales. Then the time series are partitioned using the PAM algorithm applied to the extracted features. The WAC procedure is useful when shifts in global and local characteristics of the time series are expected;
4. characteristic-based clustering (CHC) procedure. A global measure describing the time series is obtained by applying summary indices about trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity (Wang, Smith and Hyndman, 2006). The normalized indices are the inputs of the PAM algorithm.

In QSC the quantile sheet has been estimated setting $K = 19$ quantile values, with probabilities $\tau_k = 0.05 + 0.05(k-1)$, $k = 1, \ldots, K$, $L = 6$ and $M = 6$ B-spline functions, with knots equally distributed over the ranges. In CQC we have set six different quantile values with six B-splines basis functions for each penalized regression. In FCC we have considered six B-spline basis functions. Smoothing parameters are selected by cross-validation pooling the time series in a unique cluster.

In evaluating the performances of the four procedure we have assumed that the number of clusters

was known. The clustering procedures have been compared by means of the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). This index varies between 0 and 1 and more closer to 1 is ARI, the better is the correspondence between the clustering and the true partition.

5.1 Example 1

Time series are generated according these models:

Cluster 1 $Y_1(t) = 2\,\mu(t) + \kappa\sqrt{\mu(t)}\,\varepsilon_1(t)$;

Cluster 2 $Y_2(t) = \mu(t) + \kappa\sqrt{\mu(t)}\,\varepsilon_2(t)$;

Cluster 3 $Y_3(t) = \varepsilon_3(t)$, $\varepsilon_3(t) \sim \mathcal{G}(\mu(t)/\kappa^2, \kappa^2)$ where $\kappa^2 = 0.5$ and $\mathcal{G}(a,b)$ is the Gamma distribution with mean ab and variance ab^2.

Non-stationarity arises from the trend component

$$\mu(t) = 1 + t/132 + \exp\left\{-(t/132 - 0.6)^2/0.05\right\}.$$

In order to asses the robustness of the procedure in presence of temporal correlation, we have generated $\varepsilon_i(t)$, $i = 1, 2$ from a multiplicative seasonal ARMA model

$$\varepsilon_i(t) = e_i(t) + \theta e_i(t-1) + \Theta e_i(t-12) + \Theta\theta e_i(t-13) \tag{11}$$

Here $e_1(t)$ and $e_2(t)$ are i.i.d. random variables such that $e_1(t) \sim \mathcal{N}(0, 1/(1+\theta^2+\theta^2\Theta^2+\Theta^2))$ and $e_2(t) \sim 3^{-1/2}(1+\theta^2+\theta^2\Theta^2+\Theta^2)^{-1}\mathcal{T}_3$, where $\mathcal{T}_g$ is the Student distribution with g degree of freedom. We have considered four pairs of parameters (θ, Θ) namely (0,0), (0.8,0), (0,0.8),(0.8,0.8). The first pair corresponds to the case of independence.

We highlight that the time series models of clusters have the same variance function but different mean function $E(Y_1(t)) = 2\,\mu(t)$, $E(Y_2(t)) = E(Y_3(t)) = \mu(t)$. Moreover time series in Cluster 1 and 2 have symmetric marginal distributions whereas the distribution is skewed to the right in Cluster 3. Under this setup the clustering of time series seems not easy as Figure 2 shows.

Table 1 presents the mean and the standard errors of ARI that we obtained for one hundred simulations. QSC appears the best method in the independent case by producing markedly higher ARI values with respect to the other methods. Moreover all ARI values for QSC, CQC and FCC decrease in presence of serial dependence. However QSC procedure, with respect to the other ones, still has a discrete ability to classify the simulated time series (ARI > 0.5). It is worth noting that

QSC outperforms CQC in all examples. This means that avoiding crossing of the quantile curves decreases the risk of incorrect classification, due to no reliable estimates of the marginal distributions.

	$\theta = 0.0$, $\Theta = 0.0$	$\theta = 0.8$, $\Theta = 0.0$	$\theta = 0.0$, $\Theta = 0.8$	$\theta = 0.8$, $\Theta = 0.8$
QSC	0.864 (0.054)	0.667 (0.095)	0.686 (0.094)	0.536 (0.050)
CQC	0.550 (0.032)	0.527 (0.020)	0.530 (0.020)	0.522 (0.018)
FCC	0.503 (0.005)	0.502 (0.008)	0.501 (0.012)	0.500 (0.016)
WAC	0.519 (0.023)	0.854 (0.071)	0.500 (0.039)	0.766 (0.084)
CHC	0.438 (0.028)	0.405 (0.023)	0.471 (0.057)	0.417 (0.067)

TABLE 1

Example 1: mean and standard errors (in parentheses) of the ARI for evaluating the performance of clustering by the quantile sheet respect to other classical methods

5.2 Example 2

In this second example we have introduced a modulated component function, namely

$$\sigma(t) = 0.2 \times (1 + t)\cos(\pi/6\, t) + 0.03 \times t^2 \sin(\pi/6\, t)$$

that reflects a time-varying seasonal fluctuation in the marginal distribution. More precisely we have considered the following models:

Cluster 1 $Y_1(t) = \mu(t) + \sigma(t) + \kappa\, \varepsilon_1(t)$;

Cluster 2 $Y_2(t) = \mu(t) + \sigma(t) + \kappa\{\mu(t) + \sigma(t)\}\, \varepsilon_2(t)$;

Cluster 3 $Y_3(t) = \varepsilon_3(t)$ and $\varepsilon_3(t) \sim \mathcal{G}(1/\kappa^2, \kappa^2\{\mu(t) + \sigma(t)\})$;

where $\kappa = \sqrt{2.5}$. Serial dependence in Cluster 1 and 2 has been introduced by means of the autoregressive model

$$\varepsilon_i(t) = \rho\varepsilon_i(t - e) + e_i(t), \qquad i = 1, 2 \tag{12}$$

where $e_i(t)$ are i.i.d. random variables such that $e_i(t) \sim \mathcal{N}(0, 1/(1 - \rho^2))$.

In this example the time series are highly heteroscedastic. (see Figure 3). The results (see Table 2) confirm the superiority of QSC procedure for independent data ($\rho = 0.0$) and moderate correlations ($\rho = 0.2, 0.4$) and the overall robustness of the procedure in case of serial dependence (ARI > 0.89).

6. CHLOROPHYLL TYPE-A CONCENTRATION IN THE ADRIATIC SEA

The clustering procedure (QSC) outlined in Section 4 was applied to the time series of the logarithm of the Chlorophyll type-a concentration in the Adriatic Sea. Such procedure was compared to the

	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$
QSC	0.951 (0.027)	0.955 (0.024)	0.945 (0.027)	0.894 (0.033)
CQC	0.707 (0.081)	0.477 (0.158)	0.301 (0.087)	0.217 (0.060)
FCC	0.042 (0.025)	0.027 (0.020)	0.023 (0.016)	0.023 (0.017)
WAC	0.487 (0.028)	0.510 (0.041)	0.720 (0.085)	0.926 (0.032)
CHC	0.536 (0.008)	0.538 (0.010)	0.540 (0.011)	0.528 (0.017)

TABLE 2

Example 2: mean and standard errors (in parentheses) of the ARI for evaluating the performance of clustering by the quantile sheet respect to other classical methods

FCC, WAC and CHC procedures.

In the QSC procedure we used model (7) with 6 B-spline functions both for the temporal trend and for each modulated component ($L_1 = L_2 = L_3 = 6$). Moreover we had chosen the probabilities $\tau_k = 0.05 + 0.05(k-1)$, $k = 1, \ldots, K$ and $M = 4$ B-spline functions. The knots of the B-spline functions were equally distributed over the ranges.

In our dataset about 90% of the time series are quite far from the coast and exhibits a similar temporal pattern so the use of a common degree of smoothing seems justified for the data we handle.

The GACV criterion selects $\lambda = 2.522 \times 10^{-3}$ and $\tilde{\lambda} = 6.656 \times 10^{-1}$ as smoothing parameters. The whole procedure has been implemented using the R language. All calculations have been carried out on a 2.40 GHz 32 core-processor with 96 GB of RAM. Using 10 cores and R-package `parallel` one evaluation of the GCV, i.e the estimation of 2168 quantile sheets, requires approximately 180 seconds.

Similarly to QSC, in FCC procedure we have considered a seasonal modulation model (Eilers et al., 2008) for each time series, namely

$$Y(t) = \sum_{l=1}^{L_1} \gamma_{l,1} B_l(t) + \sum_{l=1}^{L_2} \gamma_{l,2} B_l^C(t) + \sum_{l=1}^{L_3} \gamma_{l,3} B_l^S(t) + \varepsilon(t) \tag{13}$$

with $L_1 = L_2 = L_3 = 6$. We have accounted for possible serial correlation in the error term $\varepsilon(t)$ by supposing that the error term follows a moving average model of order 6. This order has been identified after preliminary analysis in which we have estimated the coefficients in model (13) using ordinary least squares.

Finally, the number of clusters for all the procedures is chosen according to the Gap-statistic, with $B = 500$ simulated reference data sets.

With the first attempt to cluster the sites, we consider the time series without regard of the spatial information.

In Figure 4 we present the spatial distribution of the clusters that we have obtained. Sites are labelled with a colour ranging from grey to green in accordance with the average value of each cluster. High values of Chl-a concentrations are associated to the dark colours (green colours in the online version of the paper).

All clustering procedures point out that there exists a value scale according the distance of the sea sites with respect to the coast. However in spite of this similarity there are several substantial spatial differences in that pattern. First of all the Gap statistic suggests a different number of clusters for each procedure. In FCC and CHC the sites are clustered into 12 clusters, while 6 clusters are required by QSC and WAC procedures. A closer analysis of the results allows us to highlight the differences in the spatial patterns (see Figure 4).

- according to FCC procedure (Figure 4-a), sites in the Northern Adriatic Sea are grouped in three clusters due to the influence of the rivers of the Po valley on the Chl-a concentration. In particular, Cluster 1 and 2, identified by dark green colours in the figure, are related to the influence area of the Po river freshwater discharges. The subsequent clusters (from 4 to 9) regroup the sites near to the Intermediate Italian coastal area and the Croatian coasts which are affected by local seasonal rivers; the last three clusters (10, 11 and 12) correspond to the open Central and Southern Adriatic Sea. This part of the basin is not significantly affected by land based inputs of Nitrogen and Phosphorus compounds. However, the overall spatial pattern appears a little blurred and is inconsistent in some areas;
- for QSC procedure the Gap statistic suggests six clusters (Figure 4-b) and in comparison with FCC results, it seems better able to resolve the spatial structure in the Northern part of the basin. The clusters are spatially more homogeneous and related to the coast and the river mouth distances. The clustering confirms the presence of a large area (Figure 4-b; Cluster 6) distant from the coast and unaltered by any potential source of anthropic influence;
- for WAC procedure the gap statistics suggests the same number of the clusters of QSC. However the spatial pattern is very different (Figure 4-c). The main difference is the presence of a large area (Cluster 2) which corresponds to the entire Northern Adriatic Sea. Cluster 1 regroups the sites near the upper Intermediate Italian coastal area, whereas Cluster 3 embraces the coastal zone of Southern Adriatic Sea and the Albanian coasts. Last three clusters correspond to zones

far from points of influence, but they have a fragmented spatial distribution;

- the clustering results obtained by CHC procedure are less convincing (see Figure 4-d). Most part of the clusters is patched with a lack of consistency.

In short, FCC and QSC procedures yield to a more consistent clustering because they clearly separate the area influenced by the plume of the Po river from the surrounding. However, QSC better resolves the spatial structure in the northern part of the basin. In fact, the spatial structure suggested by FCC is more fragmented and less consistent with general knowledge about the evolution of Chl-a in the basin. In fact the regrouped sites in the central and the southern western coastal area don't follow a clear gradient when we move away from the coastline. Furthermore, in the northern part, the area directly affected by the plume of Po river is not clearly identified.

In the previous analysis we do not take into account the spatial information. The maps of the clusters can be perturbed by local factor resulting from variance instability and outliers in the raw time series so sites that are very far apart could possibly be assigned to the same cluster with high probability. As a consequence, similarity between time series far apart could be stronger than that of observations near each other, which runs counter to cluster structures often desired in spatial setting.

For the QSC procedure we have corrected the entries of the dissimilarity matrix in the PAM algorithm using (10). The variogram model $\Gamma(h;\theta)$ has been identified by visual inspection of the graphical representation of the moment estimates (9) of the semi-variogram against the distances (see Figure 5). The moment estimates have been calculated using the coefficients of the B-spline functions which have been detrended by means of a first order linear surface trend. The Figure 5 suggests that an exponential model with a nugget effect, namely

$$\Gamma(h;\theta) = \theta_1(1 - \exp\{-h/\theta_2\}) + \theta_3(1 - \mathbf{1}_0(h)) \qquad \theta_1, \theta_2, \theta_3 > 0.$$

fits quite well.

Mutatis mutandis we can apply the same idea for correcting the entries of the dissimilarity in the FCC and WAC procedure as in Giraldo, Delicado and Mateu (2012); Haggarty, Miller and Scott (2015). Regrettably, we cannot extend such corrections to CHC method because it is not clear how defining an appropriate distance measure between the proposed finite set of time series characteristics which can be compared to the L2 norm used in the other procedures.

If we compare the obtained practical ranges (see Figure 5), i.e. the distances where the semi-variogram first reaches 95% of the sill, the spatial smoothing effects is more evident for QSC than FCC and WAC.

The Gap statistic, applied to the spatially weighted dissimilarity matrix, points out new numbers of clusters with respect to the previous results (see Figure 6). For FCC procedure this number decreases from 12 to 10 clusters: the grouping is slightly coarse and the results appear too simplified, in particular if we consider the areas near to the Italian coast. For WAC procedure the number of clusters is reduced to 5 and the spatial distribution of the sites appears quite smoothed but too simplified.

Instead the number of clusters increases for QSC procedure from 6 to 12. The resulting clusters seems to be able to better catch the influence of the freshwaters from the rivers defining a unique cluster (Cluster 1) closed to the Po river's mouth. In addition, looking to the southern part of the Adriatic Sea, QSC permits to define a more detailed classification of the Albanian coastal waters in contrast to the FCC that exhibits less clear distinctions in such area.

In Figure 7 we show the temporal behaviour of the logarithm of Chl-a concentrations in each cluster identified by the spatially adjusted QSC procedure. The clusters are sorted by ascending average value of the time series. For each cluster, we represent the quantile curves corresponding to the probabilities 0.05,0.25,0.50,0.75 and 0.95. Such curves are obtained by fitting a quantile sheet using the time series that belong to the cluster.

Sites belonging to Cluster 1 are located near to the Po river's mouth which is the most important influence point of this basin for Chl-a concentration. They exhibit the highest average value with respect to all the other clusters. The temporal evolution of the grouped time series shows an increasing trend in the period 2010-2012; the seasonal component is weak and more pronounced in the last years. Cluster 2 collects sites in front of the upper Intermediate Italian coasts. It presents a seasonal pattern (the so-called "Algae bloom cluster") with the highest peaks during the beginning of years 2011 and 2012. The temporal trend of Cluster 3 appears more similar to Cluster 1 with the seasonal peaks barely visible and high values in the last part of the temporal window.
Covering the extreme Northern Adriatic Sea, sites of Cluster 4 are spatially comprised between sites of Cluster 2 and the Dalmatian coast. We can observe a strong seasonality, more evident than the

previous clusters. Cluster 5 and 6 encompasses several areas disseminated near the coasts of the Southern Italy and the Albany. Despite the spatial distribution of these clusters interests the same areas, they show a different temporal pattern: Cluster 5 is more similar to the "coastal" Cluster 1 and 3, while Cluster 6 is characterized by a strong seasonality likewise the Cluster 4 (but with lower values). Cluster 7 and 8 regroup the sites placed between the coastal zone and the offshore areas: these sites are not directly affected by rivers and other points of influence, but their trend point out a halfway behaviour between "bloom" and "coastal" clusters. It is interesting to consider that sites belonging to Cluster 9 are closed to the Dalmatian coasts, but its temporal trend is completely different with respect to the other coastal clusters (1, 3 and 5), with low Chl-a concentrations and a moderate seasonality.

Finally, sites of Cluster 10, 11 and 12 are located far from the coasts, from north to south. As consequence of this location, the sites belonging to these clusters report lower Chl-a values and a prominent seasonal pattern, due to the solar radiation and the temperature ("no bloom cluster").

We finally remark that our findings are consistent with previous attempts of classifying Adriatic waters on the basis of Ocean Colour data. For example, D'Ortenzio and Ribera d'Alcalà (2009), using a K-means procedure, identified four main clusters within the Adriatic Sea: 2 different coastal clusters with a high seasonal signal, characteristic of nutrient enriched coastal waters; an intermittently blooming areas in the southern part; a "no bloom" cluster in the remaining zones (the eastern part of the Adriatic Sea). In the QSC clustering, Cluster 10, 11 and 12 regroup the typical evolution of Adriatic oligotrophic areas. As one can see from Figure 7, which shows the estimated quantile sheets for each cluster, this area is characterized by low Chl-a values all year around and it is impossible to detect clear multi-annual trends. Cluster 1, 3 and 5 overlap with the "coastal water" clusters described in (D'Ortenzio and Ribera d'Alcalà, 2009), but our clustering allows us to better resolve the spatial gradient and to clearly identify the influence of the Po river on Chlorophyll type-a dynamics. In fact, the clearly visible increasing trend in the years 2008-2012, which may be due to higher river discharges of inorganic nutrients which, in turn, caused a higher primary production. These features are more marked in Cluster 1, which is directly affected by the plume of the Po River. In addition the Adriatic marine current circulation transports the nutrients of the Po river to south and the clusters follow the distance from this point of "source" (Giani et al., 2012). The absence of

important rivers explains the difference of classification inherent to the eastern coast of the Adriatic Sea with respect to the Croatian coast. A separate discussion concerns the Albanian coastal area which reports a temporal pattern similar to the Italian coast (Marini et al., 2010).

7. DISCUSSION

In this work we have proposed a new procedure to cluster time series of satellite data especially suitable when we deal with marginal distributions that vary over time and have missing values. We have applied the procedure to GlobColour data related to Chlorophyll type-a concentrations in order to identify homogeneous areas in the Adriatic Sea with respect to the temporal behaviour of this water indicator. In the literature there exist several attempts to cluster water quality indicators (Henderson, 2006; Pastres, Pastore and Tonellato, 2011; Haggarty et al., 2012) that consider time series under a functional data analysis perspective. A distinctive feature of our proposal is that we use the information coming from the quantiles of a time series. Such information is conveyed in terms of the quantile sheet of a time series, which is a surface in the time-probability space. The use of quantile sheet, not allowing the quantiles curve to cross in the fitting procedure, leads to reliable estimates of the conditional distributions. The importance of this constraint has been evident in our small simulation study. Our procedure reports the highest values of the Adjusted Rand Index with respect to a procedure that forms group using separated quantile curves. Moreover our simulation results show some degree of robustness of the procedure in the case of correlated observations, even if a wider study is undoubtedly required. Finally differently from the aforementioned references we take into account the role of the spatial dependence in the clustering procedure.

Concerning our specific water quality indicator, many studies have used the Chlorophyll type-a concentration in order to assess the spatial distribution of water eutrophication in the Adriatic Sea (D'Ortenzio and Ribera d'Alcalà, 2009; Giani et al., 2012; Djakovac et al., 2012; Marini et al., 2010; Mélin et al., 2011), however its clustering in homogeneous zones is still an open problem. The clusters derived by our procedure has showed convincing results.

It is worth pointing out that Chlorophyll type-a concentration appears to be impacted by several factors. Therefore one potential future development of this work is the multivariate extension resulting in the analysis of "quantile volumes" (Schnabel and Eilers, 2013) with the aim to classify using information from several physical attributes (Ramos et al., 2012).

Another future development is a better description of the spatial dependence. At this stage our proposal can be resumed in two steps: a summary of the temporal behaviour and a clustering procedure, corrected by spatial information. We are currently working on a model-based method for clustering random time-varying functions that are spatially interdependent combining the proposal in Jiang and Serban (2012) and in Reich (2012).

AKNOWLEDGEMENTS

We are indebted with the anonymous referees for their constructive comments that have improved highly the original version of the paper.

REFERENCES

Abraham, C., Cornillon, P. A., Matzner-Løeber, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* **30** 581–595.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing* **11** 1350003.

Behrenfeld, M. J. and Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography* **42** 1–20.

Bondell, H. D., Reich, B. J. and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97** 825–838.

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans* **100** 13237–13254.

Cheng, K. F. (1983). Nonparametric estimators for percentile regression functions. *Communications in Statistics - Theory and Methods* **12** 681–692.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, revised ed. Wiley, New York.

Djakovac, T., Degobbis, D., Supić, N. and Precali, R. (2012). Marked reduction of eutrophication pressure in the northeastern Adriatic in the period 2000–2009. *Estuarine, Coastal and Shelf Science* **115** 25–32.

D'Ortenzio, F. and Ribera d'Alcalà, M. (2009). On the trophic regimes of the Mediterranean Sea: a satellite analysis. *Biogeosciences* **6** 139–148.

Eilers, P. H. C., Currie, I. D. and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis* **50** 61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–121.

Eilers, P. H. C., Gampe, J., Marx, B. D. and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine* **27** 3430–3441.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics* **26** 78–89.

Giani, M., Djakovac, T., Degobbis, D., Cozzi, S., Solidoro, C. and Umani, S. F. (2012). Recent changes in the marine ecosystems of the northern Adriatic Sea. *Estuarine, Coastal and Shelf Science* **115** 1–13.

Giraldo, R., Delicado, P. and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* **66** 403–421.

Haggarty, R. A., Miller, C. A. and Scott, E. M. (2015). Spatially weighted functional clustering of river network data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64** 491–506.

Haggarty, R. A., Miller, C. A., Scott, E. M., Wyllie, F. and Smith, M. (2012). Functional clustering of water quality data in Scotland. *Environmetrics* **23** 685–695.

He, X. (1997). Quantile curves without rossing. *The American Statistician* **51** 186–192.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics* **17** 65–80.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9** 60–77.

Huot, Y., Babin, M., Bruyant, F., Grob, C., Twardowski, M. and Claustre, H. (2007). Does chlorophyll a provide the best index of phytoplankton biomass for primary productivity studies? *Biogeosciences discussions* **4** 707–745.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* **8** 231-255.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98** 397–408.

Jiang, H. and Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, New York.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge, UK.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition* **38** 1857 - 1874.

Marini, M., Grilli, F., Guarnieri, A., Jones, B. H., Klajic, Z., Pinardi, N. and Sanxhaku, M. (2010). Is the southeastern Adriatic Sea coastal strip an eutrophic area? *Estuarine, Coastal and Shelf Science* **88** 395–406.

Maritorena, S., d'Andon, O. H. F., Mangin, A. and Siegel, D. A. (2010). Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. *Remote Sensing of Environment* **114** 1791–1804.

Mélin, F., Vantrepotte, V., Clerici, M., D'Alimonte, D., Zibordi, G., Berthon, J.-F. and Canuti, E.

(2011). Multi-sensor satellite time series of optical properties and chlorophyll-a concentration in the Adriatic Sea. *Progress in Oceanography* **91** 229–244.

Nieto-Barajas, L. E. and Contreras-Cristán, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Analysis* **9** 147–170.

Pastres, R., Pastore, A. and Tonellato, S. F. (2011). Looking for similar patterns among monitoring stations. Venice Lagoon application. *Environmetrics* **22** 712–724.

Petitjean, F., Inglada, J. and Gançarski, P. (2012). Satellite image time series analysis under time warping. *Geoscience and Remote Sensing, IEEE Transactions on* **50** 3081–3095.

Piccolo, D. (1990). A distance measure for classifying ARMA models. *Journal of Time Series Analysis* **2** 153-163.

Ramos, E., Juanes, J. A., Galván, C., Neto, J. M., Melo, R., Pedersen, A., Scanlan, C., Wilkes, R., van den Bergh, E., Blomqvist, M., Karup, H. P., Heiber, W., Reitsma, J. M., Ximenes, M. C., Silió, A., Méndez, F. and González, B. (2012). Coastal waters classification based on physical attributes along the NE Atlantic region. An approach for rocky macroalgae potential distribution. *Estuarine, Coastal and Shelf Science* **112** 105 - 114.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61** 535-553.

Schlossmacher, E. J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association* **68** 857–859.

Schnabel, S. K. and Eilers, P. H. C. (2013). Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis* **97** 77–87.

R Core Team (2014). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B* **63** 411-423.

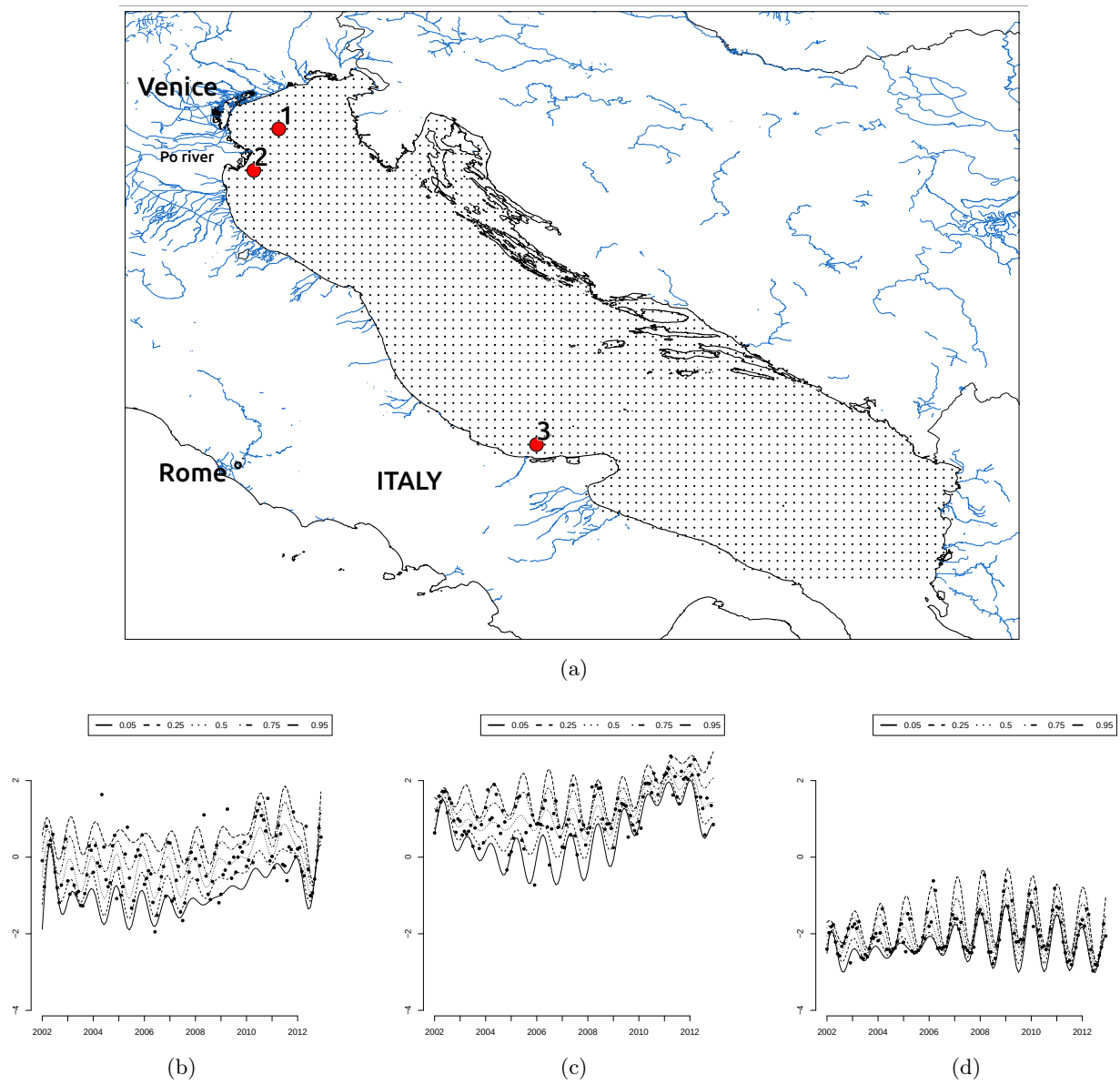
Wang, X., Smith, K. and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery* **13** 335–364.

Yoder, J. A., McClain, C. R., Feldman, G. C. and Esaias, W. E. (1993). Annual cycles of phytoplankton chlorophyll concentrations in the global ocean: a satellite view. *Global Biogeochemical Cycles* **7** 181–193.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics & Data Analysis* **50** 813–829.

Scientific Campus,
via Torino 155,
I-30170 Venezia Mestre, Italy
E-mail: gaetan@unive.it
E-mail: paolo.girardi@unive.it
E-mail: pastres@unive.it

260 Route du Pin Montard,
F-06904 Sophia-Antipolis, France
E-mail: antoine.mangin@acri-st.fr



(a) (b) (c) (d)

Fig 1: Study area (a) and logarithm of the observed values along with the estimated quantile curves at 3 sites (1 (b), 2 (c) and 3 (d), respectively) in the period 2002-2012.

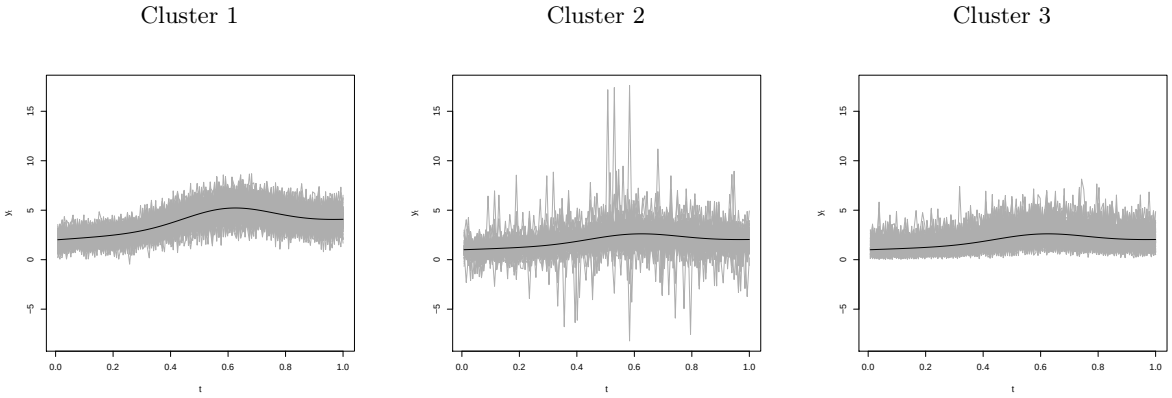


Fig 2: Example 1: one hundred simulated trajectories with $\theta = 0$ and $\Theta = 0$ in (11). Solid lines represents the mean function $E(Y_i(t))$.

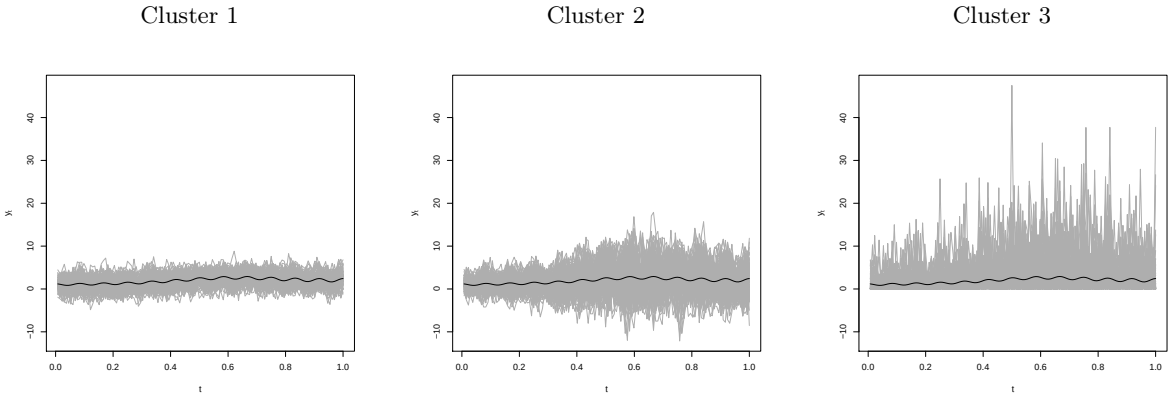


Fig 3: Example 2: one hundred simulated trajectories with $\rho = 0.6$ in (12). Solid lines represents the mean function $E(Y_i(t))$.

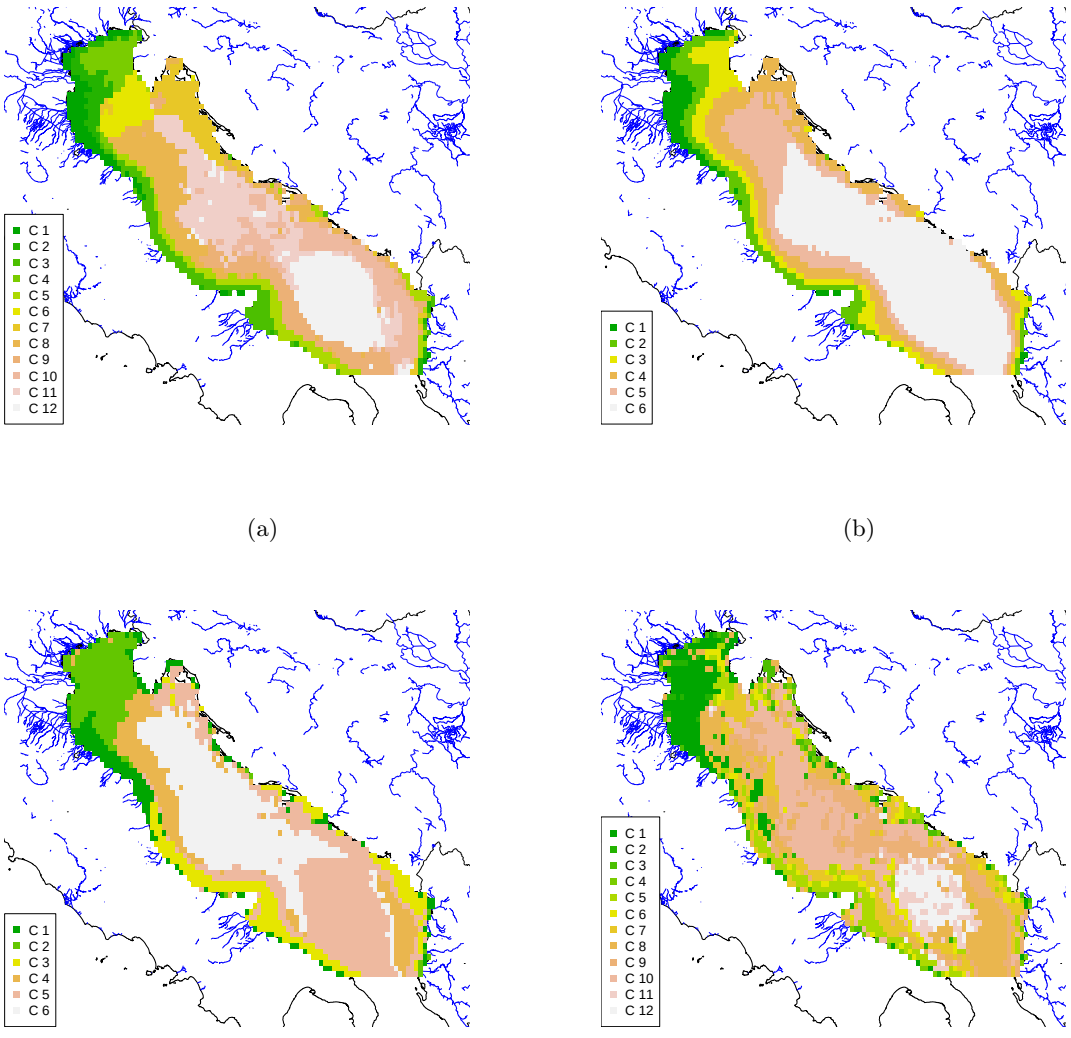
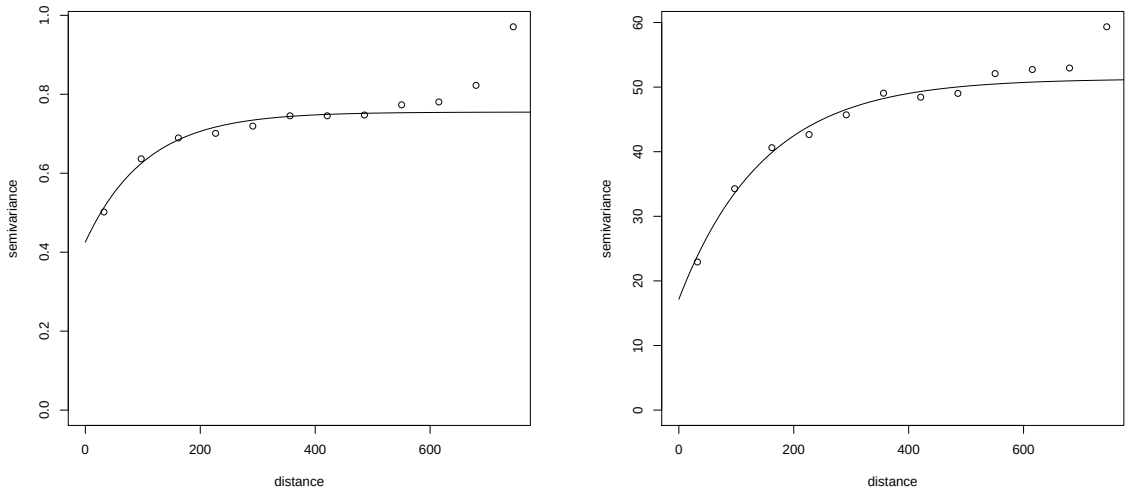
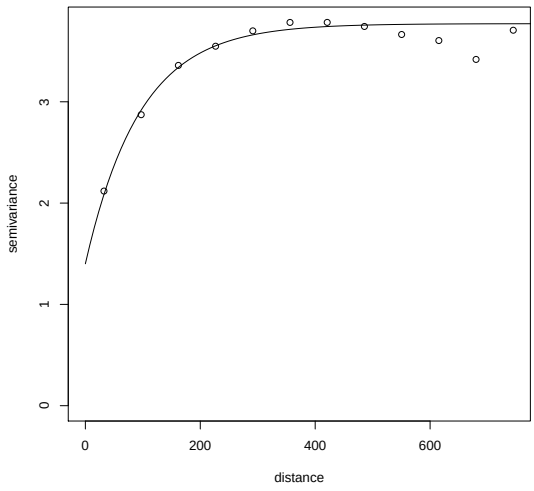


Fig 4: Spatial distribution of the clusters in the Adriatic Sea identified by (a) FCC, (b) QSC, (c) WAC and (d) CHC.



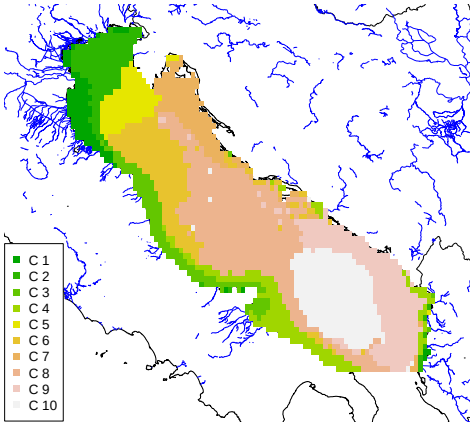
(a)

(b)

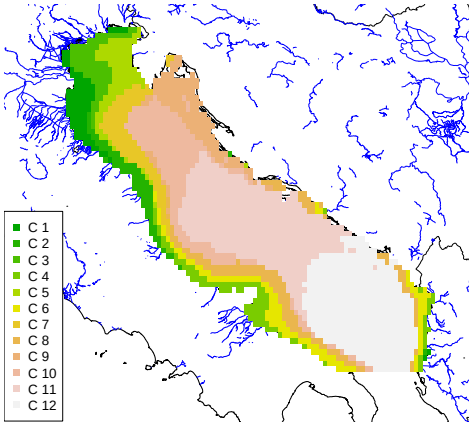


(c)

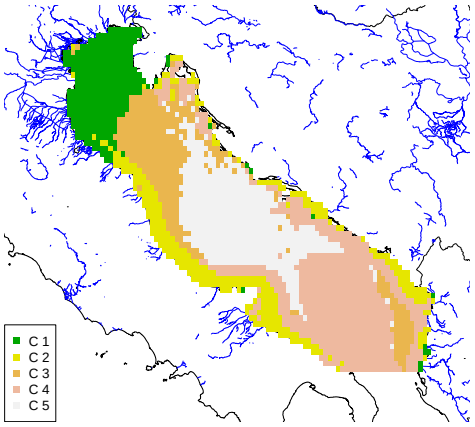
Fig 5: Empirical and fitted semi-variogram for (a) FCC, (b) QSC and (c) WAC clustering procedures. In all cases the coefficients of the B-spline functions have been detrended using a first order linear trend surface.



(a)



(b)



(c)

Fig 6: Maps of the clusters in the Adriatic Sea resulting from the clustering procedures (a) FCC, (b) QSC and (c) WAC after correcting the dissimilarity matrix by the spatial weights.

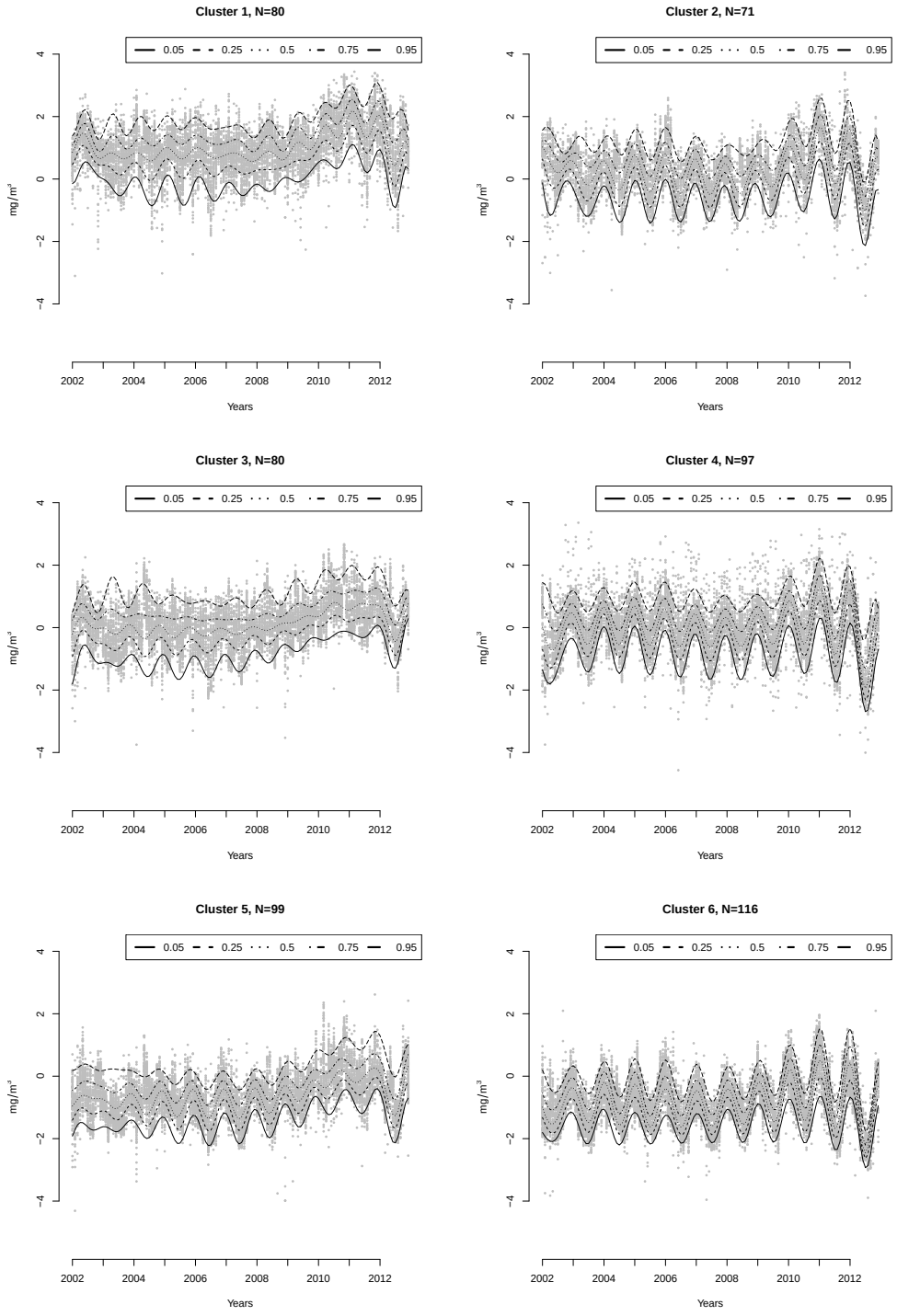


Fig 7: Chl-a distribution for the first six clusters of the Adriatic Sea with five estimated quantile curves by QSC with spatial adjustment procedure. In the title the number of time series that belong to the cluster is reported.

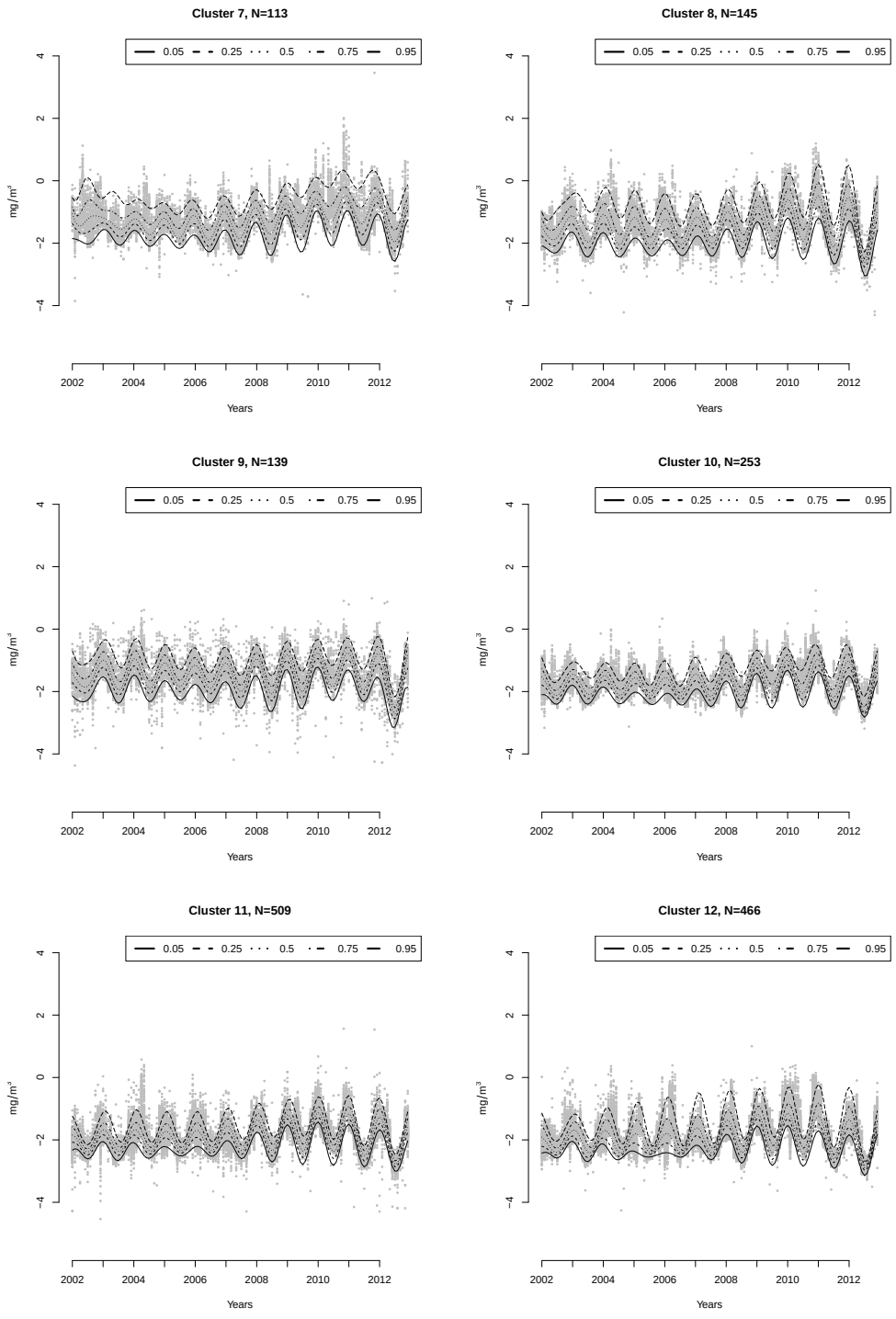


Fig 8: Chl-a distribution for the last six clusters of the Adriatic Sea with five estimated quantile curves by QSC with spatial adjustment procedure. In the title the number of time series that belong to the cluster is reported.